MuiruriVivian /
PHASE-2-GROUP-7-PROJECT

Code Issues Pull requests Actions Projects Wiki Security In

PHASE-2-GROUP-7-PROJECT / README.md

MuiruriVivian Update README.md f1214ae · 30 minutes ago

169 lines (88 loc) · 10.4 KB

Preview Code Blame Raw

GROUP-7-PROJECT

PHASE 2: Final Project Submission

Please fill out:

- Students name:

  i. **Vivian Muiruri** (vivian.muiruri@student.moringaschool.com)

  ii. **Calvin Angoye** (calvin.angoye@student.moringaschool.com)

  iii. Dominic Oseko (dominic.oseko@student.moringaschool.com)

  iv. Soudie Okwaro (soudie.okwaro@student.moringaschool.com)

  v. Winfred Karimi(winfred.karimi@student.moringaschool.com)

  vi. Anguista Kupeka (anguista.kupeka@student.moringaschool.com)

  vii.

- Student pace: self paced / part time / full time; **PART TIME**

- Scheduled project review date/time: **FRIDAY, 11th 2024**

- Instructor name: SAMUEL KARU

- Blog post URL: https://github.com/MuiruriVivian/PHASE-2-GROUP-7-PROJECT

Business Problem

Your company now sees all the big companies creating original video content and they want to get in on the fun. They have decided to create a new movie studio, but they don’t know anything about creating movies. You are charged with exploring what types of films are currently doing the best at the box office. You must then translate those findings into actionable insights that the head of your company's new movie studio can use to help decide what type of films to create.

The Data Source

The data is in the data folder in the repo which consist of:

> IMDB (https://www.imdb.com/)

> bom.movie_gross.csv.gz

> tn.movie_budgets.csv.gz

> tmdb.movies.csv.gz

The files are in different formats that is, CSV and TSV files and can be opened using the pd.read_csv. Data from IMDB is in a SQLite Database. We combined

> tmdb.movies.csv.gz and IMDB (https://www.imdb.com/) to form **cleaned merged_data.csv**

and

> bom.movie_gross.csv.gz and tn.movie_budgets.csv.gz to form **cleaned gross_budget.csv**

Importing the Libraries

The following libraries were imported for data analysis and visualization tasks:

- numpy for high level mathematical functions, working with Arrays and performing statistical calculations **import numpy as np**
- pandas data manipulation, reading, analysis and managing data structure for tablular data **import pandas as pd**

- seaborn and matplotlib for data visualization such as bar charts, line plots, and scatter plots **import seaborn as sns** and **import matplotlib.pyplot as plt %matplotlib inline**
- matplotlib image for image upload as **import matplotlib.image as mpimg**
- Sqlite3 in Python for loading the SQLite library, which provides a lightweight, disk-based database system as **import sqlite3**
- Scikit-Learn library for classification, regression, clustering, and dimensionality reduction, along with utilities for preprocessing, model selection, and evaluation as **import sklearn as sk**
- Statsmodels library, a robust statistical package that is widely used for conducting statistical tests, modeling, and data exploration as **import statsmodels.api as sm**
- Stats module from SciPy to provides a vast range of statistical tools, including probability distributions, statistical tests, and functions for descriptive statistics, for data analysis and hypothesis testing as **import scipy.stats as stats**

Data Cleaning

The four dataset will go through sanity check first thats is data cleaning, it includes:

- Converting some columns that are are supposed to be numerical i.e production_budget, domestic_gross_y, worldwide_gross from object dtype to float dtype
- Splitting the genre column to make it easier to analyze each genre independently.
- Check the null or missing values and fill them, and drop where need be
- Check and drop duplicates
- Check and drop outliers

Objectives

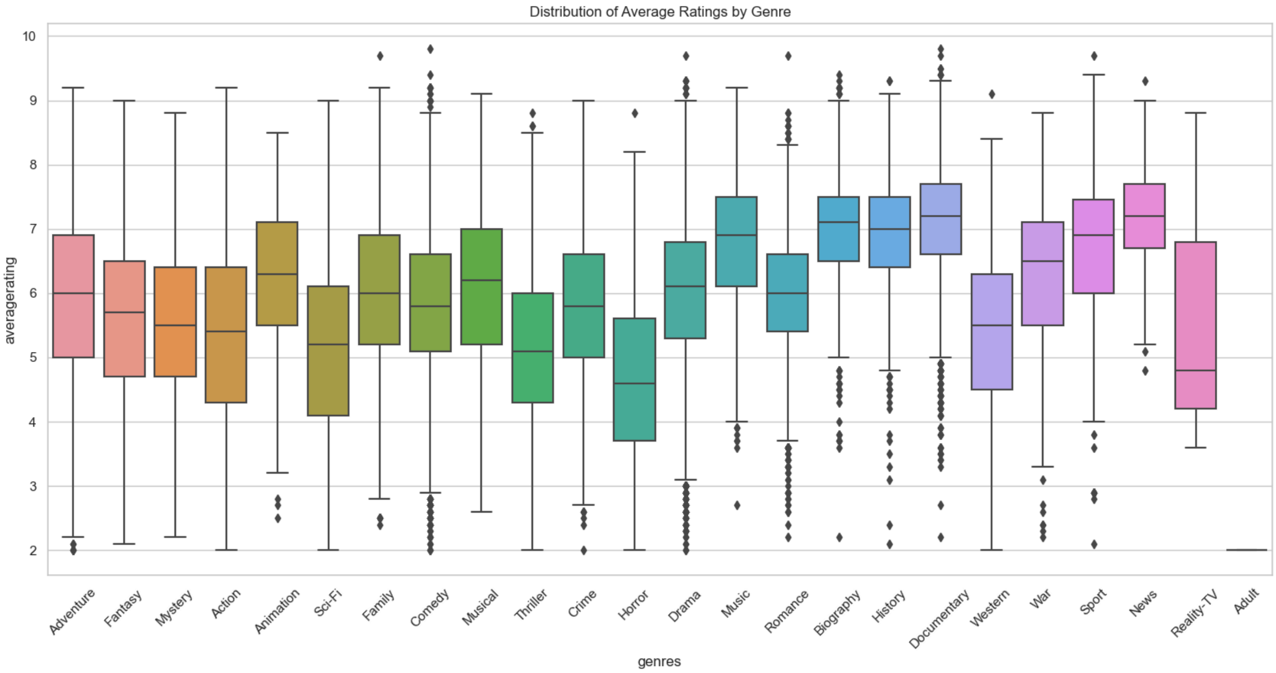
- Identify Popular Film Genres by popularity
- Identify which type of film are profitable
- Identify Emerging Trends and Audience Preferences
- identify months with highest profit

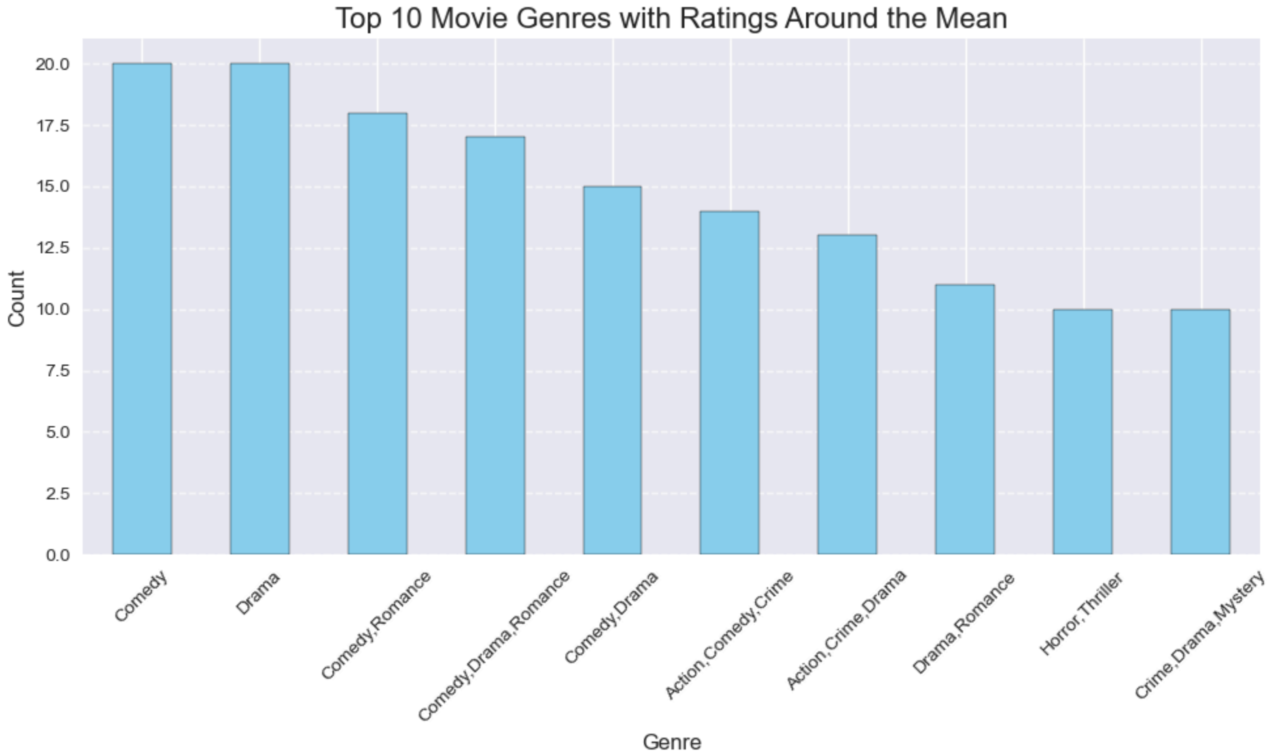
Data Analysis

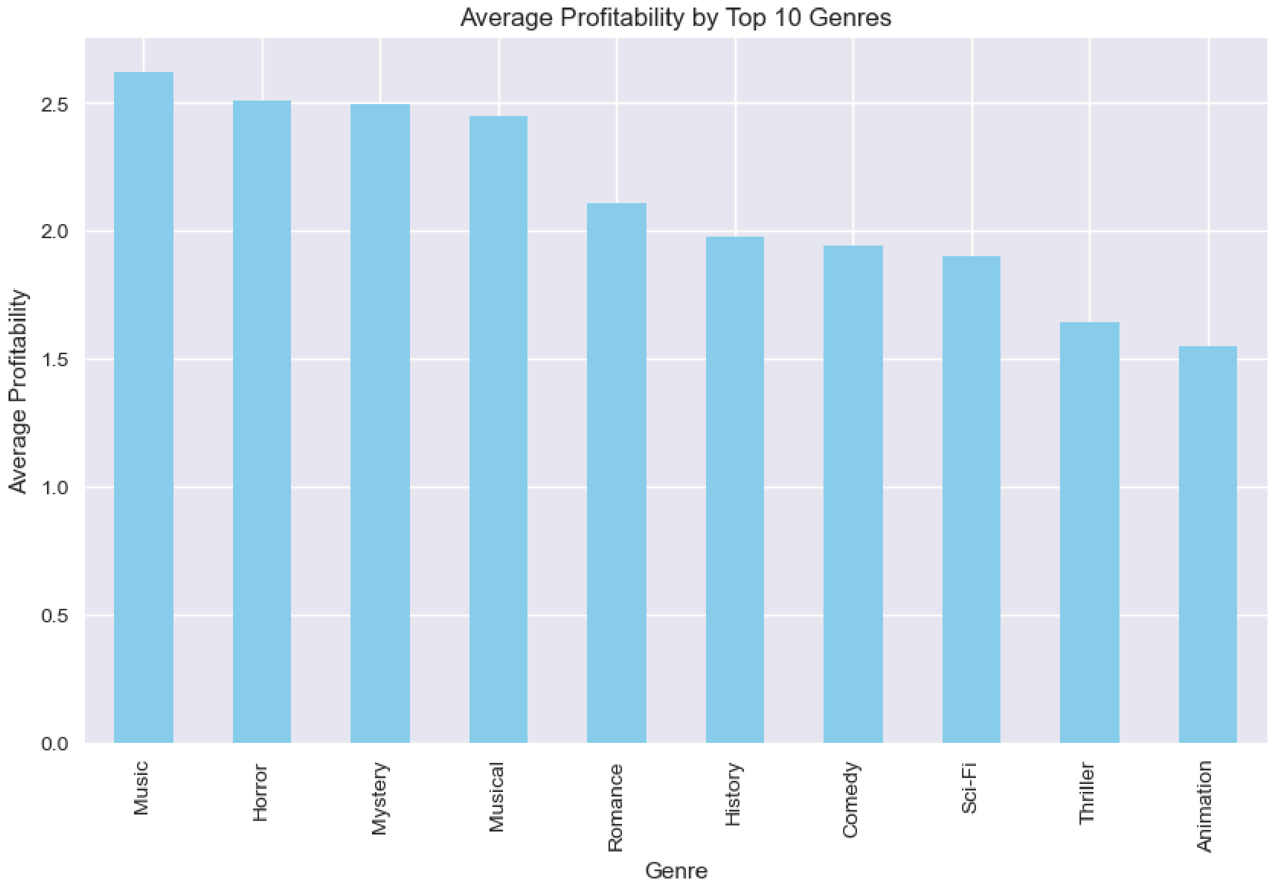
Our Data analysis structure includes:

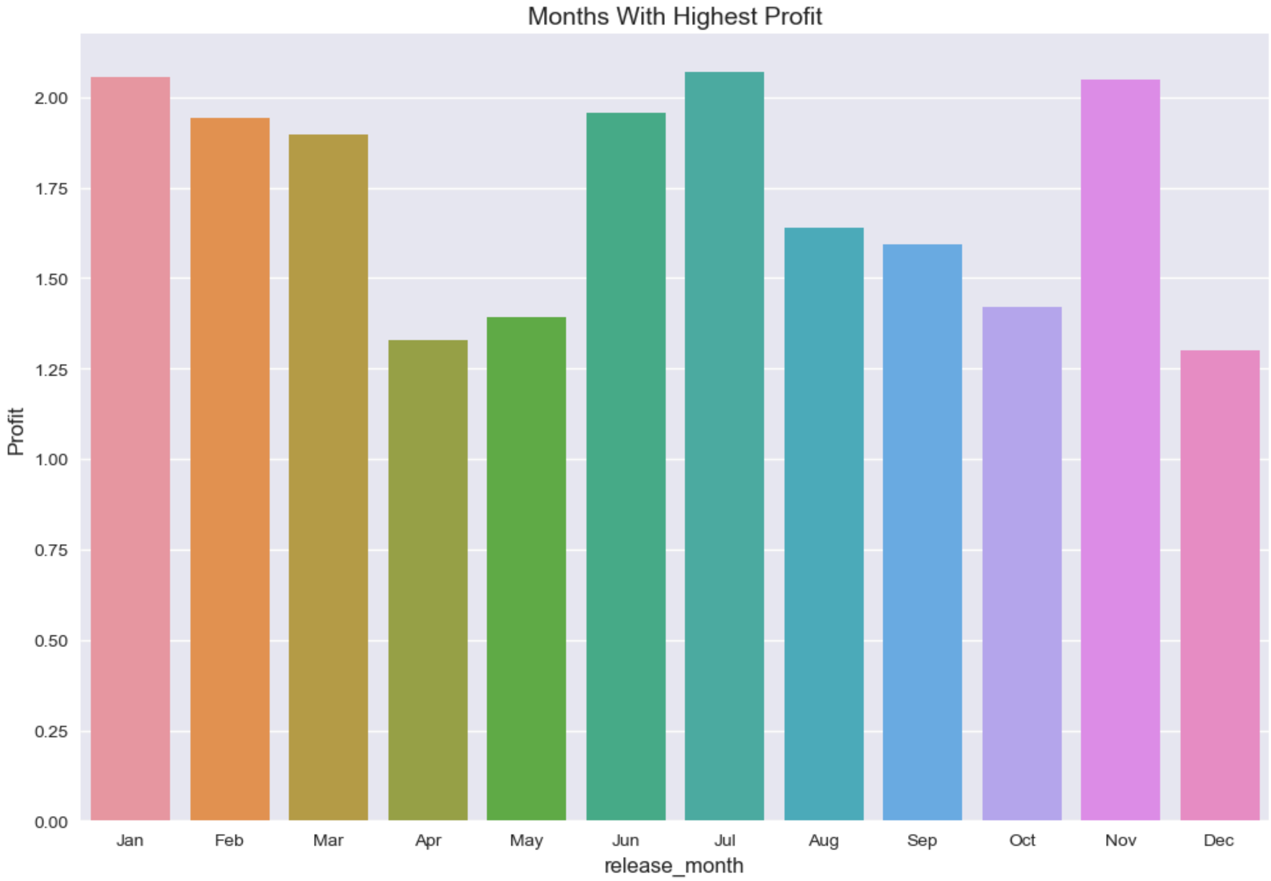
- Introduction
- Exploratory Data Analysis (EDA)
- Statistical Distribution
- Inferential Statistics
- Conclusion
- Recommendation

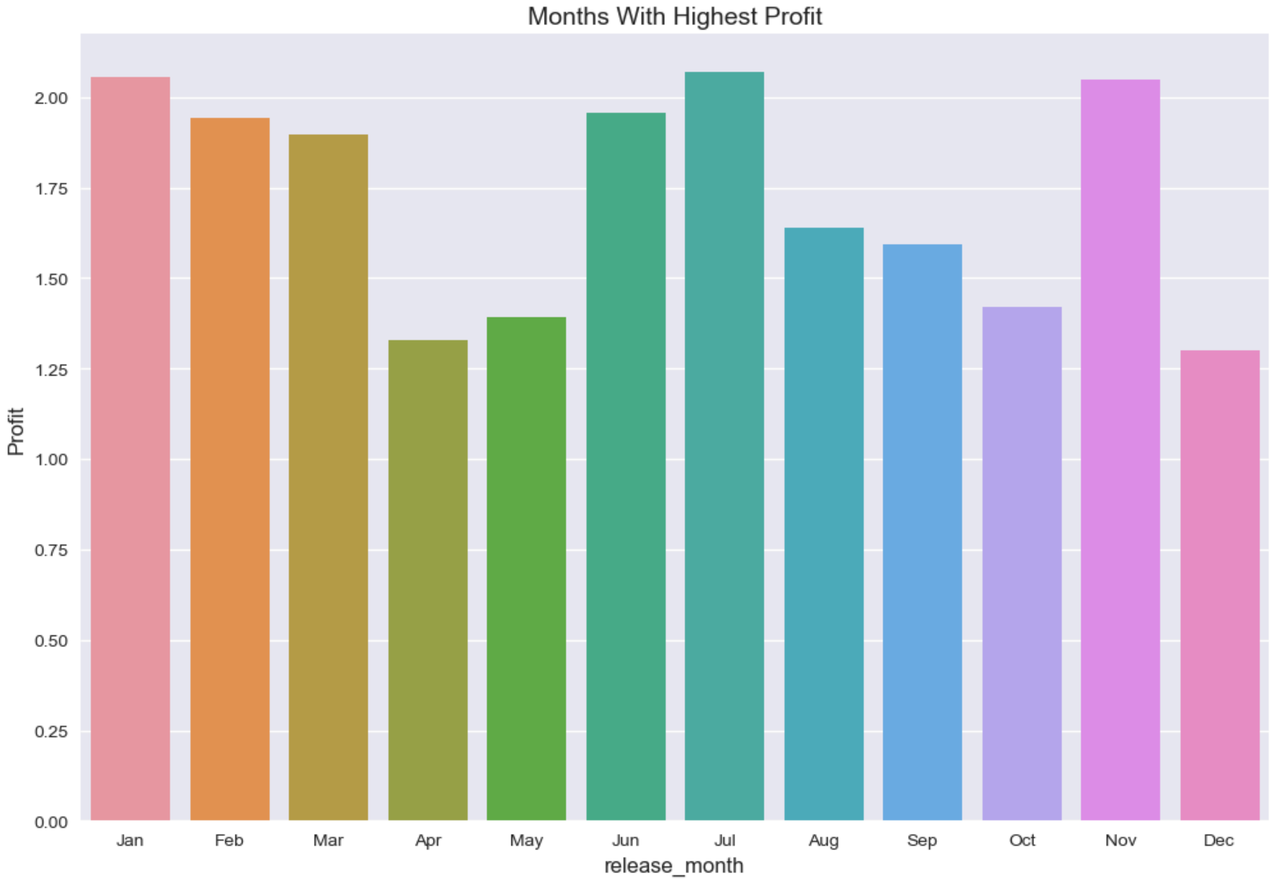
Visualizations




Top 10 Movie Genres with Ratings Around the Mean
Count
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Comedy
Drama
Comedy,Romance
Comedy,Drama,Romance
Comedy,Drama
Action,Comedy,Crime
Action,Crime,Drama
Drama,Romance
Horror,Thriller
Crime,Drama,Mystery
Genre


Average Profitability by Top 10 Genres
Average Profitability
2.5
2.0
1.5
1.0
0.5
0.0
Music
Horror
Mystery
Musical
Romance
History
Comedy
Sci-Fi
Thriller
Animation
Genre


Months With Highest Profit
Profit
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
release_month


Months With Highest Profit
Profit
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
release_month

Findings

- Most consistently rated genres are Drama and Comedy, the highest profitability is seen in Music and Horror genres, and releasing films in January, July, and November tends to maximize profits.
- The descriptive statistics reveal that the average movie has a moderate rating (mean rating of 6.36), with budgets and worldwide grosses varying widely (mean budgets around 38.6 million and gross around 90.9 million) and runtime averaging approximately 106 minutes. Profitability exhibits substantial variability, with a mean of 1.68 and a large standard deviation (2.82), indicating that while some movies are highly profitable, others experience losses, as shown by the negative minimum profitability value. The wide ranges and standard deviations across financial metrics suggest high variability in movie performance and budgets.
- The scatter plots show a weak positive relationship between runtime_minutes and averagerating, indicating that movie runtime has little effect on user ratings, while production_budget and worldwide_gross exhibit a more noticeable positive trend, suggesting that higher budgets tend to correspond with higher box office revenues.
- The data appears to be visibly normally distributed for average rating, production budget, and profitability, while runtime minutes are skewed right, with most movies falling between 90 and 120 minutes.
- The Jarque-Bera test results indicate that all the variables analyzed—averagerating, production_budget, runtime_minutes, and profitability—are not normally distributed. This conclusion is based on rejecting the null hypothesis, as the test detected significant deviations in skewness and kurtosis for all these variables. This implies that their distributions differ substantially from the bell-shaped curve of a normal distribution, suggesting potential asymmetry or heavy tails in the data.
- The histograms of the variables—production_budget, runtime_minutes, averagerating, and profitability—all show approximately bell-shaped distributions, indicating some similarity to a normal distribution.
- The correlation matrix shows a strong positive relationship between production budget and worldwide gross (0.71), while profitability is moderately positively correlated with worldwide gross and average rating, and there are weak to no correlations between the remaining variables.
- R-squared of 0.50, indicates that production_budget explains 50% of the variance in worldwide_gross, with a coefficient (slope) of 1.85 and an intercept of 19,705,776

- There is a strong correlation between a movie's financial success (worldwide gross) and both its production budget and profitability, while runtime and average rating show no significant impact. Based the attached notebook with analytics and insights, in one robust sentence, draw a general conclusion or conclusions from finding and curate at least 3 recommendations to respond to this business problem

Conclusion

The analysis indicates that movie profitability is influenced by several factors, including genre, release timing, and production budget. Genres such as Music and Horror stand out as particularly profitable, even though they often have lower budgets compared to other genres. Additionally, releasing films during certain months—specifically January, July, and November—tends to maximize box office returns, likely due to favorable seasonal demand. While higher production budgets are correlated with increased worldwide gross, profitability remains highly variable, indicating that simply spending more does not guarantee financial success. Audience ratings, measured as averagerating, show weak correlation with runtime, suggesting that longer movies do not necessarily receive better ratings. Overall, financial success appears to be more closely tied to strategic budget allocation, genre selection, and release timing rather than factors like runtime or minor increases in user ratings.

Recommendation

1. Prioritize Genre Selection for Profitability: Focus on Music and Horror genres, which show high profitability potential. These genres often resonate with niche audiences and can achieve strong box office performance without the need for excessive production budgets. This approach allows the studio to tap into reliable revenue streams while managing costs effectively.

2. Implement Seasonal Release Strategy: Schedule film releases during January, July, and November to optimize profitability by aligning with periods of higher consumer interest and lower competition in the box office. Tailoring release schedules to these strategic windows can help new releases capture a larger share of audience attention and boost revenue potential.

3. Allocate Production Budgets Based on Expected ROI: While higher budgets can drive worldwide gross, focus on optimizing budget according to each film’s potential return on investment (ROI), particularly for high-grossing genres. Avoid excessive spending on films where high budgets may not significantly enhance profitability. Instead, prioritize efficient budget use by carefully assessing the target audience, expected revenue, and genre-specific spending norms.

4. Explore Marketing and Audience Engagement Strategies for High-Return Genres: Since profitability varies widely, strengthen marketing strategies tailored to Music and Horror audiences. By effectively engaging fans through targeted advertising and promotional campaigns, the studio can maximize the visibility and appeal of these genres, enhancing box office performance and profitability.

5. Consider Audience Preference Metrics Over Runtime or Rating Increases: Given the weak correlation between runtime and user ratings, focus less on extending runtime for the sake of ratings and more on delivering quality content that aligns with audience preferences. This strategy can prevent unnecessary production costs tied to longer runtimes and instead channel resources into other value-adding areas, such as special effects or casting that enhance the movie's appeal.